{tag}

{/tag}
International Journal of Computer Applications

Volume 95 - Number 23

Year of Publication: 2014

Authors:

Liya Thomas

Syama R

10.5120/16733-6903

{bibtex}pxc3896903.bib{/bibtex}

Abstract

MapReduce is a programming model used by Google to process large amount of data in a distributed computing environment. It is usually used to perform distributed computing on clusters of computers. Computational processing of data stored on either a file system or a database usually occurs. MapReduce takes the advantage of locality of data, processing data on or near the storage areas, thereby avoiding unnecessary data transmission. The simplicity of the programming model and the automatic handling of node failures hiding the complexity of fault tolerance make MapReduce to be used for both commercial and scientific applications. As MapReduce clusters have become popular these days, their scheduling is one of the important factor which is to be considered. In order to achieve good performance a MapReduce scheduler must avoid unnecessary data transmission. Hence different scheduling algorithms for MapReduce are necessary to provide good performance. This paper provides an overview of four different scheduling algorithms for MapReduce namely; Scheduling algorithm in Hadoop, Longest Approximate Time to End (LATE) MapReduce scheduling algorithm, Self-Adaptive MapReduce(SAMR) scheduling algorithm and Enhanced Self-Adaptive MapReduce scheduling algorithm(ESAMR). An overview of these techniques is provided through this paper. Advantages and disadvantages of these algorithms are identified.

References

- J. Dean and S. Ghemawat, "Mapreduce: simplied data processing on large clusters," (2004) in OSDI 2004: Proceedings of 6th Symposium on Operating System Design and Implemention,(New York), pp. 137–150, ACM Press.
- J. Dean and S. Ghemawat, "MapReduce: a flexible data processing tool,"(2010) Communications of the ACM, vol. 53, no. 1, pp. 72–77.
- C. Jin and R. Buyya, "MapReduce programming model for . NET-based distributed computing," (2009) in Proceedings of the 15th European Conference on Parallel Processing.
- " Apache Hadoop. " http://hadoop. apache. org.
- Hadoop Distributed File System, http://hadoop. apache. org/hdfs.
- M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica, "Improving mapreduce performance in heterogeneous environments," (2008) in 8th Usenix Symposium on Operating Systems Design and Implementation, (New York), pp. 29–42, ACM Press.
- Quan Chen; Daqiang Zhang; Minyi Guo; Qianni Deng; Song Guo; , "SAMR: A Self-adaptive MapReduce Scheduling Algorithm in Heterogeneous Environment,"(2010) Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on , vol. , no. , pp. 2736-2743.
- Xiaoyu Sun, Chen He and Ying Lu "ESAMR: An Enhanced Self-Adaptive MapReduce Scheduling Algorithm"(2012) IEEE 18th International Conference on Parallel and Distributed Systems.
- R. Nanduri, N. Maheshwari, A. Reddyraja, and V. Varma, "Job aware scheduling algorithm for mapreduce framework,"(2011) in Proceedings of the 3rd International Conference on Cloud Computing Technology and Science, CLOUDCOM '11, (Washington, DC, USA), pp. 724–729, IEEE Computer Society.
- "K-means. " http://en. wikipedia. org/wiki/K-means clustering.
- G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings,"(2002) in Proceedings of the 11th international conference on Information and knowledge management, CIKM '02, (New York, NY, USA), pp. 600–607, ACM.

Index Terms

Computer Science Algorithms

Keywords

MapReduce; Programming model; Scheduling algorithms;